Lecture 8
Naive Bayes

KMA Solaiman

Fall 2023

Partially Adapted from
Tom Mitchell

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's:

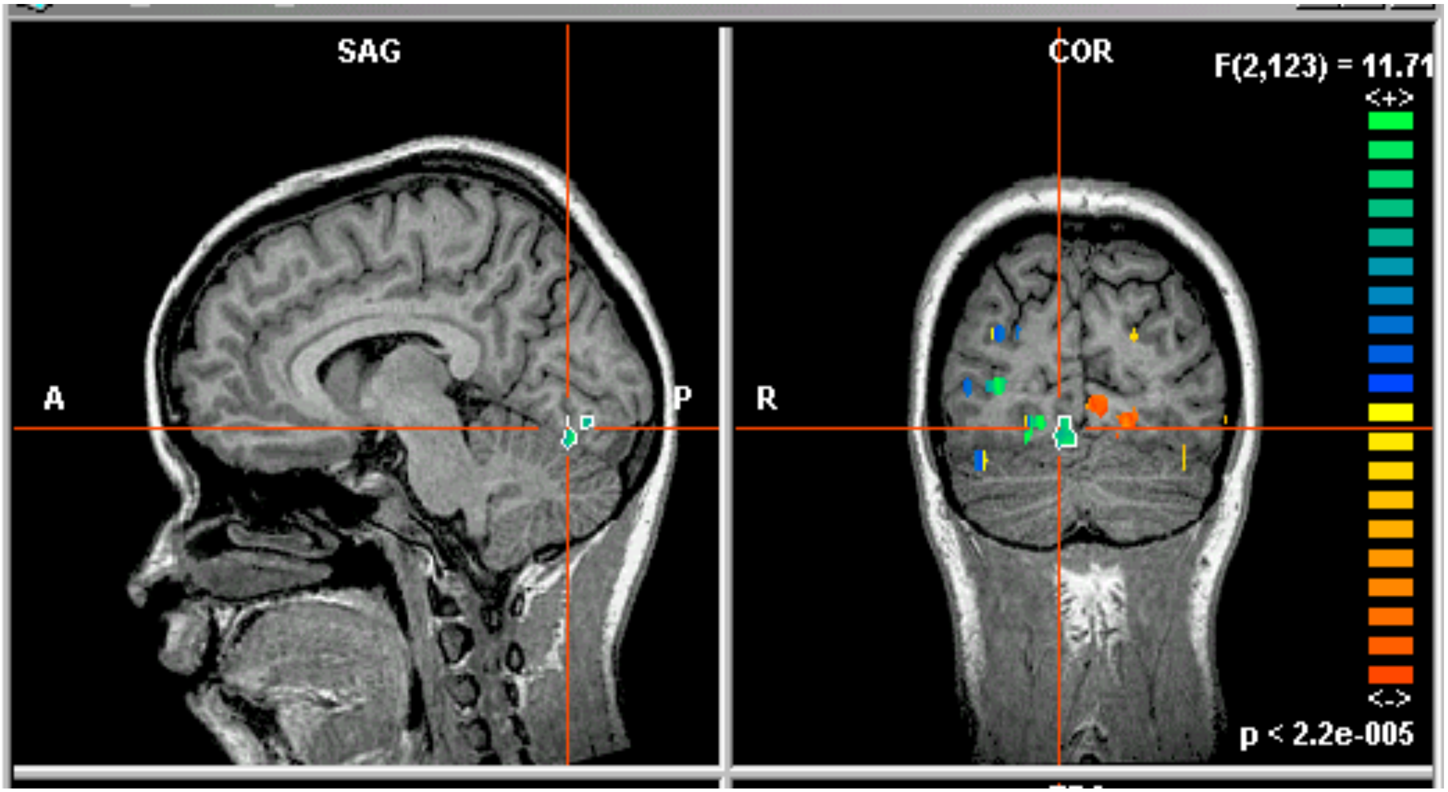
$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, classification rule for $X^{new} = < X_1, \ldots, X_n >$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

What if we have continuous X_i ?

Eg., image classification: X_i is real-valued i[th] pixel



What if we have continuous X_i ?

Eg., image classification: X_i is real-valued ith pixel

Naïve Bayes requires $P(X_i \mid Y=y_k)$, but X_i is real (continuous)

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution

What if we have continuous X_i ?

Eg., image classification: X_i is real-valued i[th] pixel

Naïve Bayes requires $P(X_i \mid Y=y_k)$, but X_i is real (continuous)

$$P(Y = y_k|X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i|Y = y_j)}$$

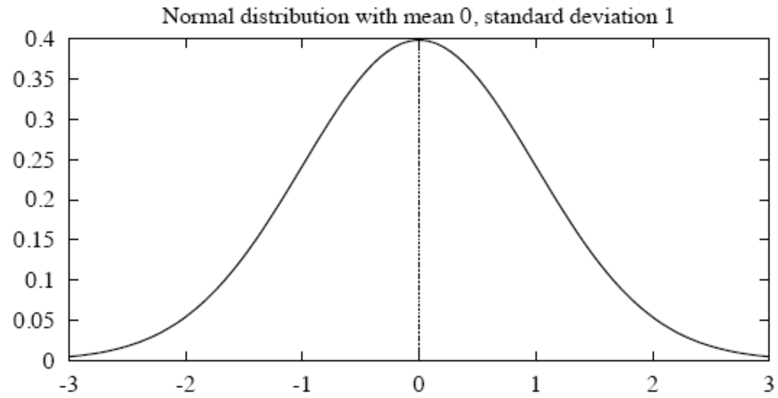
Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution

Y still follows Bernouli Distribution

Gaussian Distribution

(also called “Normal”)

p(x) is a *probability density function*, whose integral (not sum) is 1



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is

$$Var(X) = \sigma^2$$

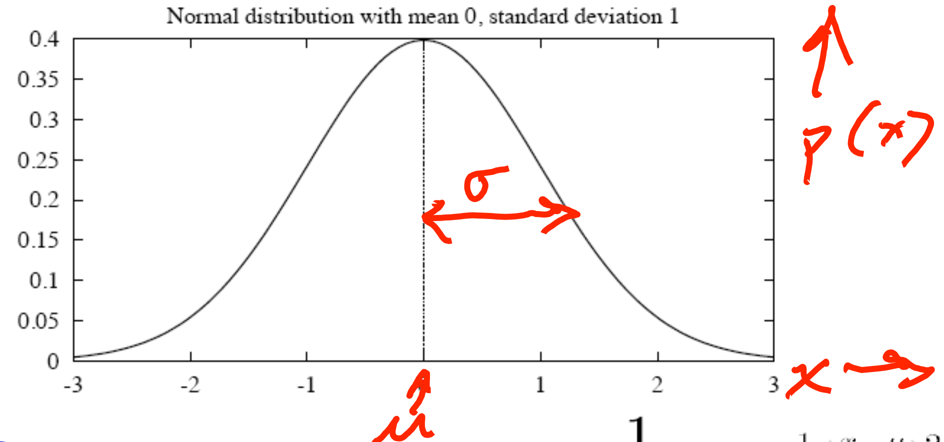
- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

Gaussian Distribution
(also called "Normal")

p(x) is a *probability density function*, whose integral (not sum) is 1

$$N(\mu, \sigma^2)$$



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is

$$Var(X) = \sigma^2$$

- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

What if we have continuous X_i ?

Gaussian Naïve Bayes (GNB): assume

$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Gaussian Naïve Bayes Algorithm – continuous X_i (but still discrete Y)

- Train Naïve Bayes (examples)

  for each value y_k

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i estimate $P(X_i | Y = y_k)$

  - class conditional mean μ_{ik}, variance σ_{ik}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \mathcal{N}(X_i^{new}; \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\widehat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

ith feature

kth class

δ()=1 if ($Y^j=y_k$) else 0

$$\widehat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \widehat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\widehat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

ith feature

kth class

$\delta()$=1 if (Y^j=y_k) else 0

$\text{Mean} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Where:

- n is the number of data points in the class.
- x_i represents the feature values for each data point within that class.

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

ith feature

kth class

δ()=1 if ($Y^j=y_k$)

```
# Sample data for two classes (continuous feature X)
class_1_data = np.array([2.1, 3.5, 1.2, 4.8, 2.9])
class_2_data = np.array([6.3, 5.7, 7.2, 5.0, 6.8])

# Calculate the class-conditional means for feature X
mean_class_1_X = np.mean(class_1_data)
mean_class_2_X = np.mean(class_2_data)
```

How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, X=<X1, … Xn>?

$$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, X=<X1, … Xn>?

$$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

Mean (μ) and Variance (σ^2) for Each Feature for Each Class: For each class, you need to estimate the mean (μ) and variance (σ^2) for each of the n features. So, for each class, there are 2n parameters to estimate (n means and n variances).

How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, X=<X1, … Xn>?

$$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

Mean (μ) and Variance (σ^2) for Each Feature for Each Class: For each class, you need to estimate the mean (μ) and variance (σ^2) for each of the n features. So, for each class, there are 2n parameters to estimate (n means and n variances).

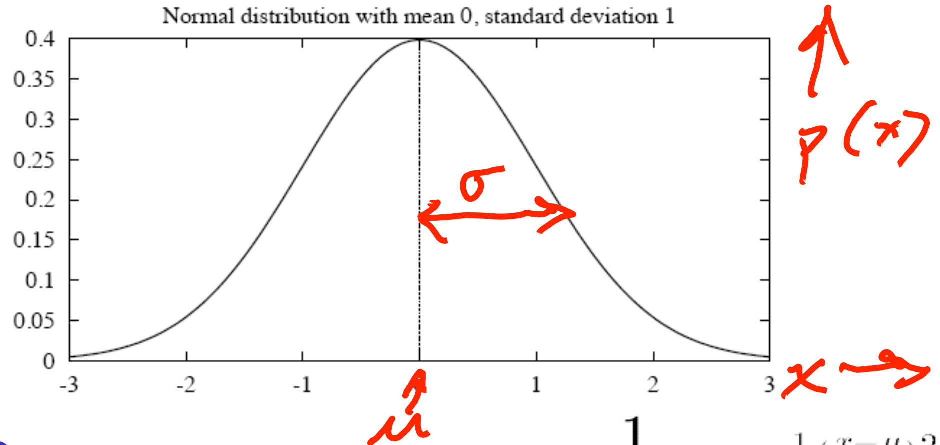
Class Prior Probability (P(Y = y)): You need to estimate one parameter for each class. So, there are k parameters to estimate.

Total Parameters = k (Class Priors) + k * 2n (Means and Variances)

Gaussian Distribution
(also called "Normal")

p(x) is a *probability density function*, whose integral (not sum) is 1

$$N(\mu, \sigma^2)$$



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

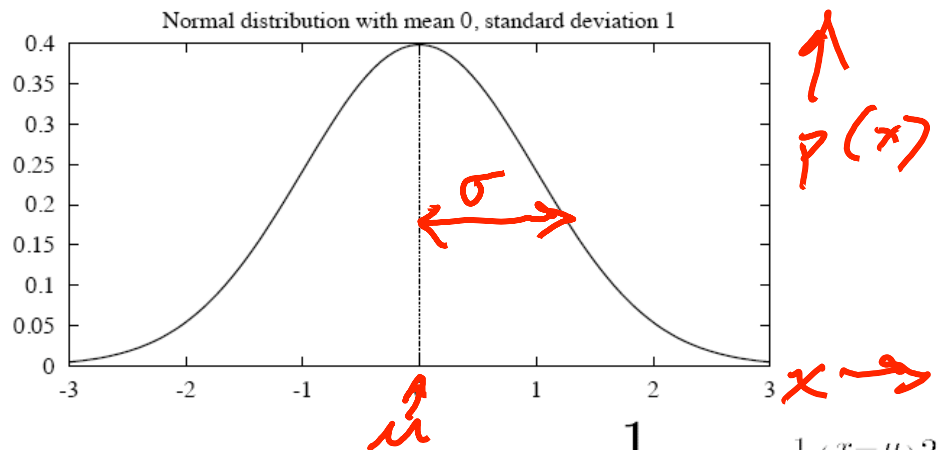
$$E[X] = \mu$$

Multivariate Normal Distribution

$$\text{“}\mathcal{N}(\mu, \Sigma)\text{”}$$

Univariate vs Bivariate

Multivariate Normal Distribution

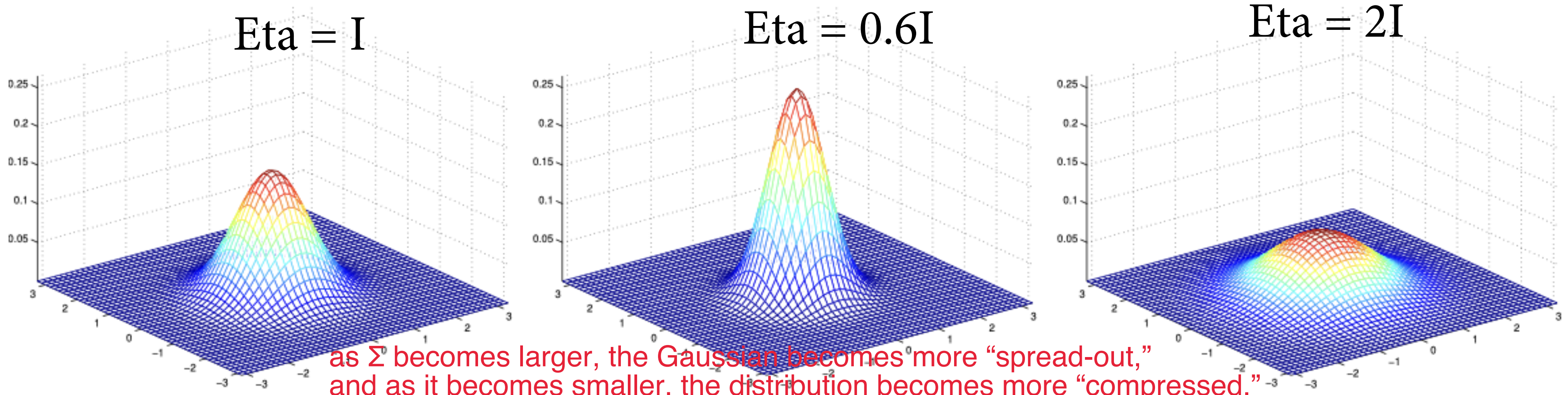


$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that X will fall into the interval (a, b) is given by

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

$$\mathrm{Cov}(X) = \Sigma.$$



as Σ becomes larger, the Gaussian becomes more “spread-out,”
and as it becomes smaller, the distribution becomes more “compressed.”

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

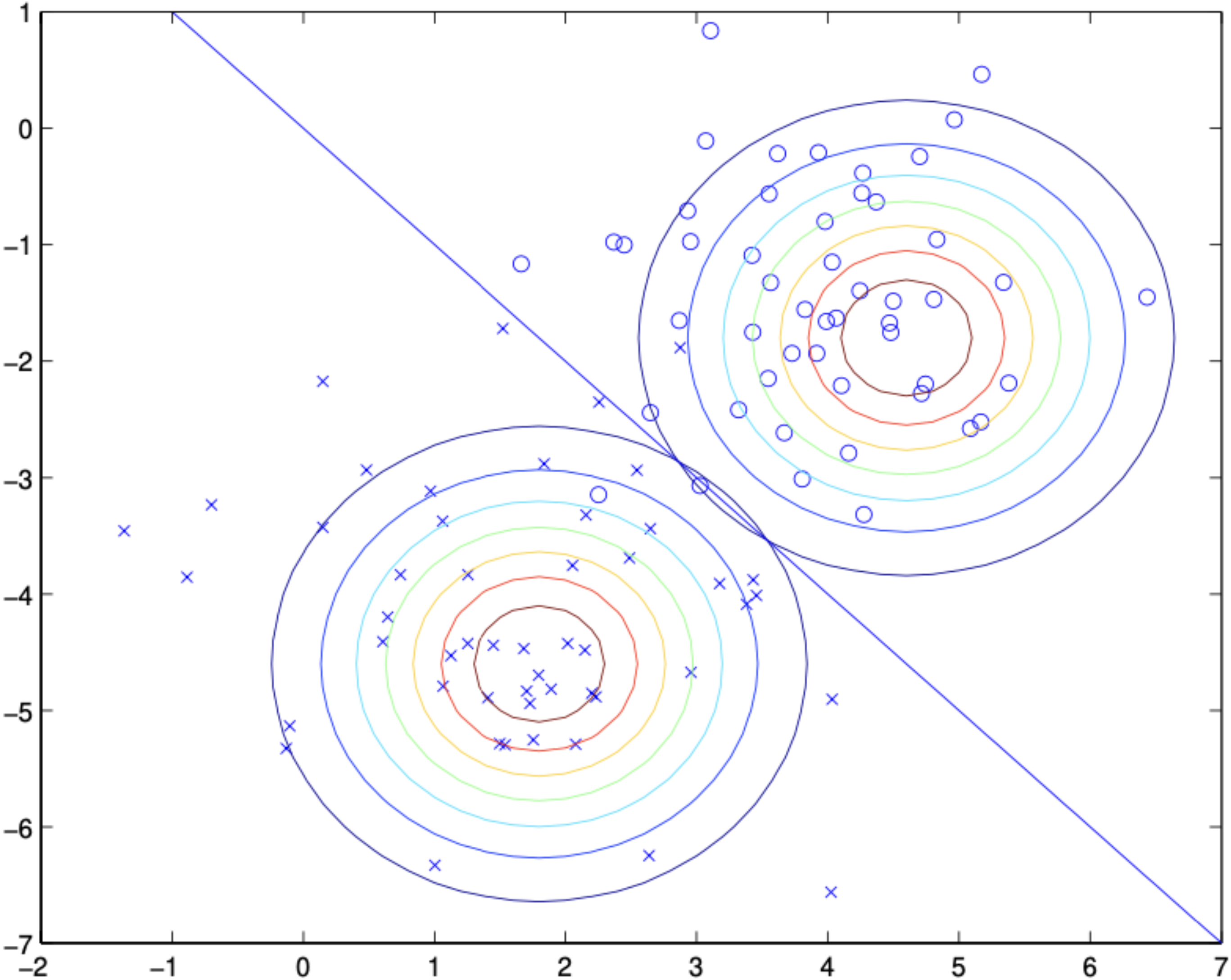
ith feature

kth class

$\delta()$=1 if (Y^j=y_k) else 0

$$\cancel{\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)}$$

Calculate Co-variance instead , Σ


1
0
−1
−2
−3
−4
−5
−6
−7
−2
−1
0
1
2
3
4
5
6
7

GDA and Logistic Regression

- if p(x|y) is multivariate gaussian (with shared Σ), then p(y|x) necessarily follows a logistic function.
- Opposite is not true
- Hence GDA makes stronger assumption
- GDA makes stronger modeling assumptions, and is more data efficient when the **modeling assumptions are correct or at least approximately correct.**
- Logistic regression makes weaker assumptions, and is significantly more robust to deviations from modeling assumptions.
- In practice, logistic regression is used more often than GDA.

$$p(y=1|x;\phi,\Sigma,\mu_0,\mu_1)=\frac{1}{1+\exp(-\theta^T x)}$$